Journal of the Korean Society of Marine Engineering, Vol. 38, No. 4 pp. 444~450, 2014 ISSN 2234-7925 Print
http://dx.doi.org/10.5916/jkosme.2014.38.4.444 ISSN 2234-8352 Online

가변 크기 문맥과 거리가중치를 이용한 동형이의어 중의성 해소

이현아†

(원고접수일 : 2014년 1월 27일, 원고수정일 : 2014년 3월 26일, 심사완료일 : 2014년 5월 8일)

Word sense disambiguation using dynamic sized context and distance weighting

Hyun Ah Lee†

요약: 의미 중의성 해소를 위한 대부분의 기존 연구에서는 문장의 특성에 상관없이 고정적인 크기의 문맥을 사용해 왔다. 본 논문에서는 중의성 해소에서 문장에 따라 가변적인 크기의 문맥을 사용하는 가변길이 윈도우와 단어간 거리를 사용한 의미분석 방법을 제안한다. 세종코퍼스의 형태의미분석 말뭉치로 학습하여 12단어 32,735문장에 대해 실험한 결과에서 제안된 방법이 용언에 대하여 92.2%의 평균 정확도를 보여 고정 크기의 문맥을 사용한 경우에 비해 향상된 결과를 보였다.
주제어: 동형이의어, 의미적 중의성 해소, 가변 크기 문맥, 거리가중치, 세종코퍼스

Abstract: Most researches on word sense disambiguation have used static sized context regardless of sentence patterns. This paper proposes to use dynamic sized context considering sentence patterns and distance between words for word sense disambiguation. We evaluated our system 12 words in 32,735sentences with Sejong POS and sense tagged corpus, and dynamic sized context showed 92.2% average accuracy for predicates, which is better than accuracy of static sized context.
Keywords: Homograph, Sense disambiguation, Dynamic sized context, Distance weight, Sejong corpus

1. 서 론

자연언어처리에서 어휘 중의성 해소(word sense disambiguation)는 주어진 문장에서 단어의 올바른 의미를 파악하는 과정으로, 검색, 번역, 분류 등 관련 시스템의 성능 향상에 영향을 미치는 중요한 문제이다. 대부분의 중의성 해소 연구는 다의어(polysemy)가 아닌 동형이의어(homograph)를 대상으로 하고 있다. 현대 표준 한국어에서 동형이의어가 차지하는 비율은 약 30%로 상당한 부분을 차지한다**[1]**. 사람들은 동형이의어에 대해 큰 불편을 느끼지 못하는데, 이는 문맥(context) 즉 문장 내에 발생하는 연어(co-occurrence)정보들을 통해 자연스럽게 의미를 파악할 수 있기 때문이다. 예를 들어 “배에 타서 맛있는 배를 많이 먹었더니 배가 아팠다”와 같은 문장에서 사람들은 인접한 내용어(content word)들을 통해 쉽게 의미를 파악할 수 있다.

의미 중의성 연구에서도 문맥을 활용한다. 특히 통계 기반 연구에서는 문맥 내에 발생한 단어들을 학습하여 이를 중의성 해소 과정에 사용한다. 따라서 올바른 문맥의 설정은 곧 시스템의 성능과 연결되는데, 대부분의 통계 기반 의미 분별 연구에서는 고정된 크기의 문맥을 사용해 왔다. 본 논문에서는 대상 단어의 좌우 문맥에서 의미 관계를 가질 수 있는 어절들에 대한 휴리스틱을 이용하여, 기존에 사용하던 고정 문맥 크기를 주어진 문장에 맞게 동적으로 사용하는 가변길이 윈도우를 제안한다.

2. 기존 연구

동형이의어에 대한 중의성 해소는 품사 기반 해결의 단순한 방법부터 코퍼스나 사전 등의 다

† Corresponding Author: Department of Computer Software Engineering, Kumoh National Institute of Technology, 16 Daehak-ro, Gumi, Gyeongbuk, 730-701, Korea, E-mail: halee@kumoh.ac.kr Tel: 054-478-7546

양한 지식에 기반한 방법에 기반한다**[2]**.

같은 품사를 가지는 동형이의어의 중의성 해소 방법은 언어자원에 기반한 방식과 지도학습, 비지도학습, 준지도학습 방식 등의 통계기반 방식, 두 가지를 결합한 방식으로 나뉜다. 사전이나 시소러스(thesaurus)와 같은 언어자원에 기반한 방식은 사전의 뜻풀이나 워드넷(WordNet)과 같은 언어자원으로부터 의미정보를 수집하여 중의성을 해소하는 방법으로, 말뭉치로부터 얻기 어려운 저빈도어의 정보를 얻기 용이하지만 변화하는 언어특성을 반영하기 어렵다. 말뭉치에 기반한 통계기반 의미분별에서는 의미태그가 부착된 말뭉치를 사용하는 지도학습이 비지도학습에 비해 우수한 성능을 보인다. 우리말의 경우 세종말뭉치의 의미태그 말뭉치 구축 이후 지도학습 기반 의미분별 연구가 활발하게 이루어지고 있다. 준지도학습은 정확도라는 지도학습의 장점과 자료 획득이 용이한 비지도학습의 장점을 결합한 모델로써 국외 중의성 해소 연구에 많이 사용되고 있다**[3]**.

우리말에 대한 중의성 해소 연구는 지식 기반 방법과 의미부착 말뭉치와 지식을 결합한 방법로 나뉜다**[4]-[10]**. **[4]**는 사전 뜻풀이에 기반하여 의미 계층구조를 생성하여 의미분별에서의 자료부족 문제를 해소하고자 하였으나, 의미 계층을 형성하기 위한 과정이 수작업에 의존적인 문제점이 있다. **[5]**와 **[6]**는 사전 뜻풀이에서 의미정보를 구축하고 의미별 사전확률(prior probability)과 거리가중치에 기반한 의미 분별 모델을 제안하였으나, 동형이의어는 다양한 패턴의 문장 속에서 발생할 수 있기 때문에 구문 특성을 반영해 개선할 여지가 있다. **[7]**은 시소러스, 기계번역사전, 세종전자사전, 말뭉치 등의 언어자원을 결합하여 온톨로지를 구축하여 중의성을 해소하였다. **[8]**은 사전 뜻풀이의 특성을 이용하여 의미정보를 정제하고 U-WIN의 시소러스를 이용하여 자료부족 문제를 해소하는 방법을 제안하였으나, 시소러스 자체의 획득 문제를 가진다. **[9]**은 대량의 의미 분석 말뭉치를 획득하기 어렵다는 점을 바탕으로 한국어 어휘의미망(KorLex)과 세종 형태분석 말뭉치를 이용한 중의성 해소방법을 제안하였다. **[10]**은 표준국어대사전에 명시된 목적격, 부사격, 보격 조사 정보를 통해 생성 확률과 전이 확률을 계산하여, HMM 모델을 통해 동사의 중의성을 해소하였다. 동사의 경우 격 정보가 동사의 의미 결정에 큰 영향을 미치지만, 사전으로부터 격 정보를 얻기가 쉽지 않으며 격 정보가 중복되어 발생하는 경우가 있다. **[11]**에서는 학습에서 사용할 수 있는 문맥에 대한 통계적 자질 선정에 대한 연구가 이루어졌다.

이와 같이 중의성 해소 연구에서는 문맥 정보를 이용하여 의미 분별을 시도한다. 인간의 경우 좁은 영역의 연접 단어들만으로도 의미적 모호성을 해소할 수 있으며**[4]**, 동형이의어를 중심으로 좌우 문맥을 설정하는 방식은 인간이 동형이의어의 의미를 분별하는 방법과 유사하다. 우리말에 대한 의미분별 연구에서는 추가적인 지식의 사용이나 통계적 기법에 의존적이었으며, 문맥을 사용하는 영어권의 연구에서는 **[12]**과 같이 자질 정보 추출에 있어서 대부분 고정된 크기의 문맥을 사용하거나 **[13]**와 같이 특정한 명사 개수를 활용하여, 문맥 크기에 대한 다양한 시도는 이루어지지 않고 있다.

3. 동형이의어 중의성 해소

본 장에서는 중의성 해소에서 고려하는 문맥의 크기를 가변적으로 결정하기 위한 방법을 제안하고, 이를 바탕으로 한 통계적 모델을 이용하여 중의성을 해소하는 과정을 설명한다.

3.1 가변길이 윈도우

어떠한 단어가 중의성을 가질 때, 대상 단어의 의미를 결정짓는 문맥의 범위를 윈도우(window)라고 한다. 윈도우 내에 존재하는 단어들은 중의성 해소에 중요한 단서가 되며 체언 혹은 용언과 같은 내용어들을 자질(feature)로 사용하게 된다. 의미분별을 위해 사용하는 윈도우 크기가 지나치게 작을 경우 자질 정보량이 부족해 올바른 결과를 낼 수 없으며, 윈도우 크기가 지나치게 클 경우 잡음이 발생하여 중의성 해소에 도움이 되지 않는 방해되는 자질 정보가 발생할 수 있다.

우리말 중의성 해소에 사용되는 일반적인 윈도우의 크기는 좌우 5어절이다**[4][8][9]**. **[4]**에서는 대상 단어를 기준으로 인접한 어절일수록 중의성 해소에 결정적인 영향을 미치는 단어를 포함할 확률이 높으며 좌우 5어절이내에 이러한 단어가 존재할 확률이 약 97.8%에 달한다고 밝혔다. 그러나 해당 범위 이내에 발생하는 모든 어휘가 올바른 의미 결정에 긍정적인 영향을 미치는 것은 아니기 때문에 주어진 문장에 적절한 윈도우 크기를 사용하는 것은 대단히 중요하다.

본 논문에서는 임의의 문장 내에 발생한 동형이의어의 의미를 결정짓기에 적합한 윈도우의 크기를 사용하는 가변길이 윈도우(dynamic sized context)를 사용한다**[14][15]**. 가변길이 윈도우는 의미 분별의 대상이 되는 단어와 의미 관계에 있는 단어를 찾기 위한 휴리스틱으로, 대상 단어로부터 인접한 어절의 품사정보를 통해 중의성 해소에 적합한 범위로 윈도우 크기를 확장한다. 윈도우 크기를 확장해가는 과정에는 조사와 접미사, 어미 정보가 활용되며 그 기준은 **Table 1**과 같다.

동형이의어가 일반조사와 결합한 체언일 경우 서술어를 찾아 우측으로 탐색한다. 이는 주어, 목적어, 보어와 같은 문장 성분이 서술어와의 관계에서 의미를 분별하는데 필요한 정보로 사용될 수 있기 때문이다. 예를 들어 아래 예문 (1)에서 동형이의어 '물가'와 '배'와 같은 주어나 보어는 서술어 '뛰었다'를 통해 의미를 분별할 수 있다. 체언에 관형격 조사가 결합된 경우에는 해당 어절이 관형적 기능을 띄게 되어 연접한 체언을 꾸미는 역할을 하므로 체언을 찾아 우측으로 탐색한다. 예문 (2)에서 첫 번째 어절 '배의'는 '선장'을 수식하는 역할을 하고 있으며, 용언 '되다'까지 탐색하는 것은 적합하지 않다. 또한 이들 중간에 위치하는 부사어 '몹시'는 '되다'와 연관되므로 함께 생략된다.

물가가 작년에 비해 배는 뛰었다.	(1)
배의 선장이 몹시 되고 싶다.	(2)
나는 선장이 되었고, 귀화를 허가받았다	(3)

동형이의어가 용언이면서 종결어미나 연결어미가 결합된 경우나 동형이의어 체언에 용언화접미사가 결합된 경우에는 주어 또는 목적어, 보어에 해당하는 체언을 찾아 좌측으로 탐색한다. 예문 (3)에서 '되었다'와 '허가받다'가 의미 분별의 대상인 경우 각각 '선장'과 '귀화'를 통해 의미적 관계를 찾을 수 있다. 용언에 관형형 전성 어미가 결합된 경우에는, 체언에 관형격조사가 결합되었을 때와 마찬가지로, 우측으로 탐색하여 체언을 찾는다. 용언에 명사형 어미가 결합된 경우에는 해당 어절의 성격이 주어, 목적어, 보어에 해당하므로 서술어를 찾아 우측으로 탐색한다.

윈도우의 크기는 대상 단어를 기준으로 양방향으로 확장될 수 있다. 의미 분별의 대상 단어를 기준으로 좌우 인접 어절들이 대상 단어를 포함한다면 지속적으로 확장된다. 확장에서는 **Table 1**의 탐색 전략을 역방향으로 사용한다. **Table 1**을 대상 품사를 기준으로 사용하면 체언 '배'는 "어미의 배"의 경우 관형격 조사 '의'가 결합된 앞 체언 '어미'를, "축조된 배"의 경우 용언화 접미사 '된'이 결합된 앞 체언 '축조'를 의미 관계에 있는 단어로 추출할 수 있다.

예를 들어 "부른 배를 안고 이 집에 이사온 첫

Table 1: Strategies to extend context window

품사와 결합기능어		탐색 전략		
		품사	방향	예
체언	일반 조사 (이/을/는 등)	용언	→	나는 밥을 먹었다 ▶ 밥을 먹었다
	관형격 조사 (의)	체언	→	선착장에 놓여있는 배의 후미에는 ▶ 배의 후미
	용언화접미사 (하, 되)	체언	←	근처로 이사했더니 거기에는 ... ▶ 근처로 이사
용언	종결어미 (다, 네, 오 등)	체언	←	나는 밥을 먹었다 ▶ 밥을 먹었다
	연결어미 (고, 며, 어 등)	체언	←	배가 고프고 목도 마르다 ▶ 배가 고프고
	명사형 전성어미 (음, 기)	체언	→	나는 소질이 있음을 알았다 ▶ 소질이 있음
	관형사형 전성어미 (ㄴ, 던)	용언	→	여기는 나가는 문이다 ▶ 나가는 문

날”의 예문에서 ‘배’의 의미 분별을 위한 윈도우 확장 과정을 살펴보자. **Table 1**의 윈도우 확장 기준을 바탕으로 먼저 ‘배를’과 ‘안고’의 관계를 확인하여 우측 1어절을 윈도우 범위로 설정한다. 좌측을 보면 ‘부른’은 용언에 관형형 어미가 결합된 경우이며 ‘배를’과 관계가 있음을 알 수 있다. 따라서 윈도우는 ‘부른’ 까지 확장하게 되어 좌우 1어절의 윈도우 크기를 가지게 된다.

가변길이 윈도우를 이용할 경우 단일 문장 내에 발생하는 다수의 동형이의어에 대해서도 중의성을 해소시킬 수 있다. 이를테면 “배에 타서 맛있는 배를 먹었더니 배가 아팠다”와 같은 문장이 주어졌을 때, 가변길이 윈도우는 **Figure 1**과 같이 각각의 배에 대하여 ‘배에 타다’, ‘맛있는 배를 먹다’, ‘배가 아프다’와 같은 3개의 문맥을 정확히 구분해 줄 수 있다.

Figure 1: Multiple homographs and dynamic sized context

3.2 중의성 해소 모델

식 (1)은 동형이의어와 3.1에서 제안한 방식으로 얻은 윈도우내 어휘가 주어졌을 때 의미를 결정짓는다.

$$WSD(H,W) = \arg max_{h_i \in H} \prod_{w \in W} P(w|h_i) \quad (1)$$

H는 동형이의어의 전체 의미이며, h_i은 동형이의어의 i번째 어깨번호에 해당하는 의미를 나타낸다. W는 가변적으로 결정된 윈도우 내의 모든 자질단어($w_1,w_2,...,w_n$)를 가리킨다. 윈도우 내에 발생한 모든 단어들이 h_i와 함께 나타날 확률을 모두 곱한 것 중 가장 높은 확률을 갖는 h_i를 최종적인 의미로 결정한다.

윈도우 내에 존재하는 자질들이 학습단계에 미처 발생하지 않은 경우 $P(w|h_i)$가 0이 되므로, 이 경우를 위해 각 공기 범주에 1을 더하여 $P(w|h_i)$를 **식 (2)**와 같이 계산한다. $CO(w,h_i)$은 단어 w와 동형이의어의 i번째 어깨번호 h_i와의 공기빈도이다.

$$P(w|h_i) = \frac{1 + CO(w,h_i)}{\Sigma_{h_m \in H}(1 + CO(w,h_m))} \quad (2)$$

가변길이 윈도우로 주변 문맥 중 의미있는 단어만을 문맥 정보로 사용할 수 있지만, 부사어 등이 복잡하게 포함된 문장에서는 맞지 않은 단어가 문맥으로 사용될 수 있다. “서울을 대표할 수 있는 다리를 만들어 보자”의 경우 ‘다리’의 가변 윈도우 문맥인 ‘만들다’는 올바른 문맥이지만, “여덟 개의 다리가 피아노를 치는 손가락처럼”에서 ‘치다’는 올바른 문맥이 아니다. 이 문제를 해결하기 위해서 본 논문에서는 동형이의어와 가까울수록 높은 가중치를 부여하여 중의성 해소의 정확도를 향상시킨다. *dist(w,h)*가 단어 *w*와 동형이의어 *h*간 거리일 때, **식 (2)** *CO(w,h$_i$)* 대신 *1/log(dist(w,h))*를 사용하여 거리가중치가 반영된 값을 구한다. 예를 들어 “다리가 피아노를 치는”에서 ‘치다’는 ‘다리’와 2어절 간격이므로 1/log2=0.67이 발생 빈도로 쓴다.

세종 의미분석 말뭉치를 **식 (1)**과 **식 (2)**를 그대로 적용해 가중치를 계산할 경우 불균형한 의미 분포에 의한 문제가 발생할 수 있다. 세종 의미분석 말뭉치에서 동형이의어의 빈도가 300개 이상인 1,906개 어휘를 대상으로 의미 분포를 분석하였을 때, 제 1어휘가 95%이상의 비율로 사용된 경우는 1,395개 어휘로 73.19%를 차지했다. 이러한 자료의 편향성은 상당수의 동형이의어를 사용빈도가 높은 의미로 분별하여 높은 정확도를 얻는데 기여할 수도 있지만, 저빈도어의 정보량이 매우 부족하여 저빈도 의미에 대한 정확한 분별을 어렵게 할 수도 있다. **Table 2**는 말뭉치에서 단어 ‘눈’과 공기한 자질단어들의 빈도수를 나타낸다. 'snow'의 의미로 쓰인 ‘눈’과 ‘차갑다’의 공기빈도는 7에 불과하지만,

Table 2: Frequency of feature words of '눈' and normalization

자질단어	eye ‘눈’ (9,866)		snow ‘눈’ (887)	
	기본	정규화	기본	정규화
차갑다(22)	15(0.68)	16(0.16)	7(0.32)	84(0.84)
빠지다(59)	50(0.85)	54(0.33)	9(0.15)	109(0.67)
겨울(74)	16(0.22)	17(0.02)	58(0.78)	703(0.97)

'eye'의 의미로 쓰인 '눈'은 '차갑다'와의 공기빈도가 15이다. '차갑다'와 'snow' 의미의 '눈'과 연관성이 높다는 것이 인간의 직관에 가까운데도 불구하고, 말뭉치 상에서 'eye' 의미의 눈의 비율이 매우 높기 때문에 이러한 현상이 발생한다.

위와 같은 사실에 기반하여 본 논문에서는 저빈도 의미의 자질빈도를 정규화를 시도한다. 정규화 빈도(normalized frequency) CO_{NF}를 계산하는 과정은 식 (3)과 같다. 식에서는 의미 h_i의 발생 비율의 역수를 곱하여 동형이의어의 어깨번호별 문장의 개수가 모두 동일하게 빈도수를 보정한다.

$$CO_{NF}(w,h_i) = CO(w,h_i) \times \frac{\sum_{h_j \in H} freq(h_j)}{freq(h_i)} \quad (3)$$

4. 실험 및 평가

본 장에서는 동형이의어와 자질단어간의 공기빈도를 이용한 통계 모델에 가변길이 윈도우를 적용하여 중의성 해소 성능을 평가한다. 평가실험을 위해 세종 의미분석 말뭉치 768,741문장, 9,000,2604 어절을 이용하였으며, 동형이의어에 해당되는 체언과 용언에 대해 실험을 수행하였다.

고정길이 윈도우와 본 논문에서 제안한 가변길이 윈도우(VMS)에 대한 결과 비교는 **Table 4**와 같다. 3.2의 정규화를 적용한 경우이며, 윈도우 크기 2는 좌우 각각 2개를 쓴 경우다. 결과에서 용언의 경우 가변길이 윈도우를 사용한 결과가 고정 윈도우를 사용한 경우보다 높은 정확도를 보였다.

Table 5는 4개 어휘에 대한 거리가중치와 정규화 사용에 따른 성능 평가 결과를 보인다. 기본 형태의 나이브 베이지안(MNB)과 거리가중치를 사용한 나이브 베이지안(WMNB), 정규화와 거리 가중치를 사용한 나이브 베이지안(IWMNB)의 결과를 나열한다. 평균적으로는 거리가중치 나이브 베이지안이 가장 우수한 정확도를 보였나, 단어에 따라 결과의 차이가 있었다.

Table 4: Precision for window size and word type(%)

	윈도우 크기					VWS
	1	2	3	4	5	
체언	69.6	82.8	83.4	83.5	83.3	83.4
용언	86.4	90.5	89.3	89.1	89.1	92.2

Table 5: Precision of various models for 4 words

체언	눈	배	달다	붓다	평균
MNB	**94.90**	82.32	85.82	89.16	88.05
WMNB	94.36	**83.92**	**87.31**	89.16	**88.69**
IWMNB	94.10	80.71	82.03	**92.77**	87.40

Table 6: Precision of predicates for window size

대상 단어		윈도우 크기			VWS
		3	4	5	
붓다	정규화적용	97.3	92.8	92.8	94.6
	정규화미적용	88.5	87.8	88.5	89.9
달다	정규화적용	86.4	86.4	88.6	88.6
	정규화미적용	81.3	82.7	80.6	85.6
의지	정규화적용	91.4	89.1	88.2	94.2
	정규화미적용	85.4	86.2	85.4	91.1
들다	정규화적용	88.7	88.9	88.7	92.4
	정규화미적용	93.4	92.6	91.6	94.7
기원	정규화적용	83.5	84.5	82.5	83.5
	정규화미적용	89.0	86.2	84.4	89.9
배	정규화적용	80.5	82.1	83.2	81.0
	정규화미적용	85.7	87.6	83.4	86.0

Table 6은 3.2에서 제안한 정규화를 적용한 경우와 적용하지 않은 경우에 대한 단어별 정확도를 보인다. 정규화를 적용할 경우 저빈도어의 판정 성능이 향상됨을 확인할 수 있다. '붓다', '달다', '의지'와 같은 단어들은 고빈도어와 저빈도어간에 약 8:2의 의미 분포를 가지고 있으며 저빈도어의 판정 성능이 향상됨에 따라 전체적 성능이 향상되었다. '들다', '기원', '배'의 경우 제 1어휘 비율은 낮지만 동형이의어의 의미 개수가 많아 주요 어휘들을 제외한 저빈도 어휘들의 차지 비율이 각각 1% 이하로, 매우 심한 편향성을 가지고 있어 자료부족 문제나 지나친 정규화로 인한 자료왜곡이 발생할 수 있다. 이 경우 정규화 과정에서 고빈도 의미의 중의성 해소 성능이 떨어졌으며 불균형한 평가데이터로 인해 전체 성능이 감소하는 현상이 발생한 것으로 분석되었다.

Figure 3은 정규화를 적용한 경우와 적용하지 않은 결과를 비교한다. 빈도가 높은 눈1에 대한 성능은 떨어졌지만, 빈도가 낮은 눈4에 대한 성능은 크

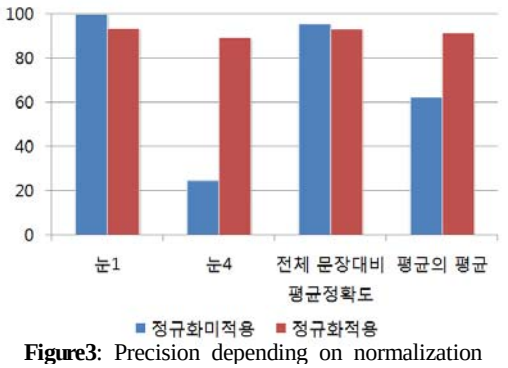

Figure3: Precision depending on normalization

게 향상되었다. 전체 평균에서는 고빈도 어휘의 비율이 높아 정규화를 사용한 경우 성능이 떨어졌지만, 단어별 정확도를 평균한 경우에는 정규화를 적용한 결과가 훨씬 우월함을 알 수 있다.

Table 7은 기존 연구의 결과와 본 논문의 결과를 비교한다. 동일한 학습 및 평가 자료가 사용된 것이 아니기 때문에 직접적인 비교가 어렵지만, 본 논문의 방식이 시소러스 등의 다양한 언어정보 없이 세종 의미분석 말뭉치만을 학습자료로 이용해 우수한 결과를 내고 있음을 알 수 있다.

Table 7: Comparison with other systems

비교 시스템	비교 어휘수	비교 시스템	제안 시스템
사전의미계층[4]	25	73.4	84.8
시소러스[8]	9	82.1	86.4
어휘의미망[9]	10	88.0	91.8
종속격정보[10]	7	91.7	92.5

4. 결 론

본 논문에서는 고정된 크기의 문맥에서 발생하는 지식부족문제와 잡음발생을 줄이고 의미 분별에 핵심적인 자질들을 추출하기 위해 가변길이 윈도우에 기반한 중의성 해소를 제안하였다. 가변길이 윈도우를 적용하였을 때의 성능을 평가하기 위해 동일한 중의성 해소 모델에서 윈도우 크기별 성능평가를 수행하였다. 용언의 경우 일반적으로 많이 사용되는 좌우 5어절의 고정 윈도우 크기에 비해 약 3.1%의 성능이 향상되었으며, 체언의 경우 고정 윈도우 크기와 비교해 큰 차이를 보이지 않아 추가적인 연구가 수행할 예정이다.

본 논문에서 제안한 가변길이 윈도우는 특정 기계학습 모델에 종속적이지 않으며, 사전이나 시소러스, 온톨로지 등의 지식을 활용한 방식에도 적용될 수 있다. 예를 들어 문서가 한정적인 분야를 대상으로 하는 경우 분야 정보를 이용하여 어휘 중의성 해소의 정확성을 높일 수 있다. 예를 들어 '선도'는 바르게 이끌어 감, 신선한 정도, 화물 인도 등의 의미를 가지는데, 해양 분야에서는 신선도의 의미로 주로 이용된다. 대상 분야 등의 지식을 제안한 방식과 함께 사용한다면 추가적인 성능향상을 얻을 수 있을 것으로 기대된다.

한국어는 교착어의 특성상 다양한 어순과 문장 패턴을 가지고 있다. 따라서 본 논문에서 제안하는 윈도우 확장 기준이 모든 유형의 문장에 대해 완벽히 처리한다고 말할 수는 없다. 그러나 본 논문에서 수행한 실험은 잡음의 제거나 가중치조절을 거치지 않고 윈도우 크기에 의해 얻은 순수한 결과로써 잡음의 제거, 가중치, 규칙정보를 활용한다면 추가적인 성능향상이 가능할 것으로 예상된다. 또한 의미 빈도에 따른 정규화 적용 방향에 대한 연구도 예정하고 있다.

후 기

이 연구는 금오공과대학교학술연구비에 의하여 지원된 논문

참고문헌

[1] B. m. Kang, “Aspects of the use of homonyms”", Language research, vol. 41 no. 1, pp. 1-29, 2005 (in Korean).

[2] J. M. Cho, Verb Sense Disambiguation Using Corpus and Dictionary, Ph.D Thesis, KAIST, 1998 (in Korean).

[3] 21st Century Sejong Project, http://www.sejong.or.kr/, Accessed May 26, 2014.

[4] J. Hur and C. Y. Ock, “A homonym disambiguation system based on semantic information extracted from dictionary definitions”, Journal of KIISE: Software and

Applications, vol. 28, no. 9, 2001 (in Korean).

[5] J. S. Kim, C. H. Kim, W. W. Lee, S. D. Lee, and C. Y. Ock, “A homonym disambiguation system based on statistical model using sense category and distance weights”, 13th Annual Conference of Human and Cognitive Language Technology, pp. 487-493, 2001 (in Korean).

[6] J. S. Kim, H. S. Choe, and C. Y. Ock, “A korean homonym disambiguation model based on statistics using weights”, Journal of KIISE: Software and Applications, vol. 30, no. 11, pp. 1112-1123, 2003 (in Korean).

[7] S. J. Kang, “Ontology construction and its application to disambiguate word senses”, The KIPS Transactions, vol. 11-B, no. 4, pp. 491-500, 2004 (in Korean).

[8] J. S. Kim and C. Y. Ock, “A korean homonym disambiguation system using refined semantic information and thesaurus”, The KIPS Transactions, vol. 12, no. 7, pp. 829-840, 2005 (in Korean).

[9] M. H. Kim and H. C. Kwon, “Word sense disambiguation using semantic relations in Korean wordnet”, Journal of KIISE: Software and Applications, vol. 38 no. 10, 2011, 554-564 (in Korean).

[10] Y. S. Par, J. C. Shin, C. Y. Ock, and H. R. Park, “Verb sense disambiguation using subordinating case information“, The KIPS Transactions, vol. 18-B, no. 4, pp. 241-248, 2011 (in Korean).

[11] Y. G. Lee, “A study on statistical feature selection with supervised learning for word sense disambiguation“, Journal of the Korean Biblia Society for Library and Information Science, vol. 25, no. 2, pp. 5-25, 2011 (in Korean).

[12] A. C. Le, A. Shimazu, V. N. Huynh, and L. M. Nguyen, “Semi-supervised learning integrated with classifier combination for word sense disambiguation”, Science Direct Computer Speech and Language, vol. 22, no. 4, pp. 330-345, 2008.

[13] D. F. Amoros, and R. Heradio, “Understanding the role of conceptual relations in Word Sense Disambiguation”, Expert Systems with Application, vol. 38, no. 8, pp. 9506-9516, 2011.

[14] G. T. Park, T. H. Lee, S. H. Hwang, B. M. Kim, H. A. Lee, and Y. S. Shin, “Korean learning assistant system with automatically extracted knowledge”, The Korea Information Processing Society Transactions on Software and Data Engineering, vol. 1, no. 2, pp. 91-102, 2012 (in Korean).

[15] G. T. Park, T. H. Lee, S. H. Hwang, and H. A. Lee, “Statistical word sense disambiguation based on using variant window size”, The 24th Annual Conference of Human and Cognitive Language Technology, pp. 40-46, 2012 (in Korean).